DOCUMENT RESUME

ED 262 559 FL 015 116

AUTHOR MacWhinney, Brian, Ed.
TITLE [Child Language Data Exchange System.]
INSTITUTION Carnegie-Mellon Univ., Pittsburgh, Pa. Dept. of Psychology.
PUB DATE Aug 84
NOTE 23p.; "Transcript Analysis" is the newsletter of the Child Language Data Exchange System (CHILDES).
PUB TYPE Guides - Non-Classroom Use (055) -- Journal Articles (080)
JOURNAL CIT Transcript Analysis; v1 n1 p2-18 Aug 1984

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Child Language; *Computer Assisted Instruction; Computer Software; *Databases; *Dial Access Information Systems; *Discourse Analysis; *Information Networks; Language Research; Microcomputers; Structural Analysis (Linguistics)
IDENTIFIERS *Child Language Data Exchange System

ABSTRACT

The background, rationale, elements, and organization of the newly-formed Child Language Data Exchange System (CHILDES) database containing child language data-transcripts are outlined in this first issue of their newsletter. The system was formed in response to a need for greater efficiency in data sharing, greater precision in data collection and analysis, and increased automation in analysis in child language studies, and consists of computerized corpora of child language. The shape of the database is described, including detailed coding information, analyses that can be performed with the available software, and limitations and problems. Appended is the first article in the second issue: "Archiving Data Sets: Issues and Problems," by Catherine E. Snow, which also provides background on the CHILDES system. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Brian MacWhinney

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED262559

TRANSCRIPT ANALYSIS

Vol. 1 No. 1: August 1984

BEST COPY AVAILABLE

Editor: Brian MacWhinney
Department of Psychology
Carnegie-Mellon University
Pittsburgh, PA 15213
Telephone: (412) 578-2656

This is the first issue of the newsletter of the Child Language Data Exchange System (ChiLDES). Comments are invited regarding the title, format, content, goals, and scope of this newsletter. Parts of the current report will be published in a forthcoming issue of the Journal of Child Language. Publication of this newsletter is supported by a grant from the John D. and Catherine T. MacArthur Foundation.

FL 015 116

Transcript Analysis

Volume 1 -- 1984

1. Editorial Comment

This is the first issue of the Child Language Data Exchange System Newsletter. It is being sent to everyone who has submitted a questionnaire indicating that they have an interest in the exchange of child language data. The current membership list may be found in chapter 5. It is my hope that it may be possible to issue this newsletter several times a year. Researchers who are interested in contributing articles to this newsletter are welcome to submit material to me at the address given on the cover sheet.[1] The newsletter is being formatted by SCRIBE and is printed on a Xerox LN01 laser printer. Issues will be structured as a series of numbered chapters. Chapters may be documents of various types: letters to the editor, member contributions, editorial contributions, appended documents, listings of the database, or program descriptions.

Members are welcome to submit requests for copies of data that are currently in the files. The current status of the database is given in chapter 6. In order to receive copies of files we need to ask members to complete the form to be found in chapter 7. Our basic membership form is given in chapter 8 and chapter 9 includes a form for those who wish to contribute data. Chapters 1 to 4 review the current status of the Child Language Data Exchange System. The Journal of Child Language has agree to publish much of the material in those chapters. Since publication will not occur for some time, it would be most helpful for members to provide comments regarding suggested changes in the material in chapters 1 to 4.

FL 015/6

[1] If at all possible, I would appreciate receiving contributions in computer-readable form. This may be done by 1) sending me floppies (Commodore, IBM-PC, or Apple), 2) sending mail to "macwhinney@cmu-psy-a" via ARPANET, BITNET, or UUCP, or 3) by logging on to a CMU machine and uploading files via Telenet.

Child Language Data Exchange System Newsletter
Dept. of Psychology, Carnegie-Mellon University
Transcript Analysis
Volume 1 -- 1984
No. 1 - August

2. Reasons for Promoting Data Exchange

The rapid expansion of child language studies during the period from the late 1950's to the present has been fueled, in large part, by the introduction of devices for audio and video taperecording. These instruments permitted the collection of large corpora of spontaneous and elicited speech of children interacting with their peers, parents, and visiting experimenters. With the introduction of the audio taperecorder and the video taperecorder, it became possible to make records of child language productions that were far more complete than the diaries that had been compiled from pencil and paper notes. With these new data, the field was able to advance well beyond the level of anecdotal summaries that had characterized the first 100 years of child language research. The new methodology permitted full recording of not just the child, but also those who interacted with the child. This new possibility slowly turned the field's attention to studies of the input to the language-learning child. During the 1960's and 1970's, researchers continually improved their use of recording technology, using directional microphones, wireless microphones, well-justified sampling strategies, non-intrusive recording methods, and detailed systems for coding aspects of the communicative context.

The proliferation of transcript data led to immense advances in specificity and accuracy of our science. It also allowed us to see more clearly the limitations involved in current analytic techniques. As we begin to compare hand-written and typewritten transcripts, problems in transcription methodology, coding schemes, and cross-investigator reliability have become more apparent. But, just as these new problems have arisen, a new major technological opportunity has emerged. This opportunity is provided by the proliferation that is now occurring of microcomputers and microcomputer software. With microcomputer word-processing systems readily available to all researchers, it is easy to enter transcript data into computer files which may then be easily duplicated, edited, and analyzed by standard data-processing techniques. The numbers of researchers entering their data in this fashion has grown exponentially over the last few years. However, it has also become apparent that this new opportunity might slip away if researchers could not move quickly to establish conventions, standardizations, and policies for the sharing and comparison of computerized transcript data.

Early in 1984, support to Carnegie-Mellon University from the John D. and Catherine T MacArthur Foundation enabled us to begin work on the establishment of a system for the exchange of such transcript data. The goal of this system is to bring about major improvements in the organizational structure and accessibility of the empirical base for child language theory. We believe that there are three basic needs to be addressed: 1) the need for greater efficiency in data sharing, 2) the need for increased precision in data collection, coding, and analysis, and 3) the need for increased automation of coding and analysis. The new systems, which we have called the Child Language Data Exchange System (ChiLDES), will utilize a variety of recent advances in data-processing hardware and software to address these three specific needs and to work toward the overgoal of improving structure and accessibility of child language data. We will explain the organization of the exchange system below, but first we must consider in greater detail the three major needs we are addressing.

2.1 The need for greater efficiency in data sharing

Child language data--transcripts of conversations between children and interlocutors--are notoriously time-consuming to produce. Moreover, any given set of transcripts is typically greatly underused and undershared. For example, taping and transcribing one child on a weekly basis requires one hour of observation, 6-10 hours of primary transcription, another 1-3 hours of transcript checking, editing, and

FL 015 116

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
X Minor changes have been made to improve reproduction quality
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

typing, all before any coding or analysis can be carried out. Such a corpus may then be used to ask questions about one aspect of language acquisition (i.e. the priority of tense vs. aspect) and then filed away on the shelf. The sad fact is that almost any corpus that is thus filed away is potentially useful for a large range of analyses and the non-availability of such data to the scientific community at large is a great loss to the field as a whole.

Because child-language data is so time-consuming to collect and to process, it is simply not feasible to undertake certain kinds of studies of great potential theoretical interest. For example, studies of individual differences in the process of language acquisition require both an intensive longitudinal analysis and large numbers of subjects - a combination which is practically impossible for a single researcher or a small research team. As a result, conclusions about differences in child language have been based on analysis of as few as two children, and rarely on groups larger than 25. However, the compilation of transcripts from many researchers would enable us to test hypotheses about patterns of relationship across various dimensions along which individuals may differ.

2.2 The need for greater precision in data collection and analysis

The reuse of transcript data for new purposes often forces the analyst to consider additional aspects of the process of data collection, transcription, coding, and analysis. Thus, the data sharing that will be promoted by the establishment of the ChiLDES is likely to lead to an acceleration in the development of interest in data collection as a process. From the point-of-view of the long-term advance of the field, this would certainly be a welcome development.

Perhaps the most important outcome of this process will be the development of new standards for data collection, transcription, and analysis. The sharing of data that have already been collected may also lead to analytic advances. With files that are currently computerized, it is relatively easy to supplement the basic text line with additional coding features, thus (a) giving richer and more complete descriptions of each data set and (b) enabling competing hypotheses to be tested against each other on a single data set. A recurrent issue in the analysis of child language data is how to determine the 'right' set of categories to use in the description; researchers have typically argued for the correctness of their own categories from their own newly collected data only, without the possibility of comparing them directly to other descriptions. For example, in the 1960's and early 1970's when semantic categories such as "agent" and "action" were used extensively by researchers, the pragmatic notions of "topic" and "comment" were not applied. However, it is perfectly feasible to go back to earlier transcripts and add categories like topic and comment to the existent semantic coding schemes, and then to compare directly the power of the pragmatic vs. the semantic categories. Moreover, whenever researchers have preserved the original audiotapes, it is possible to supplement transcripts with full phonetic and prosodic coding.

Another example of the possibilities for better explanation opened by an accretional coding system involves adding interactional features to the coding, in an attempt to explain child language characteristics. For example, Folger & Chapman (1978) concluded from their analysis of 6 children that individual differences in children's tendency to imitate may relate to parental imitativeness, and that certain parental speech acts were much more likely to be imitated by children than others. If a central data exchange system had been available to Folger & Chapman, they could have added their speech act analysis to the transcripts for the six children whose imitations had been extensively studied by Bloom, Hood, & Lightbown (1974), to test directly the hypothesis that it is the shape of parental speech acts rather than the state of the child's knowledge that predicts children's imitation.

2.3 The need for increased automation in analysis

The advent of the taperecorder led to the first automatization in the study of child language. The coming of accessible personal computing opens up the door to a second and potentially even more important automatization. With handwritten or typewritten transcripts, scanning for the occurrences of particular events, word types, syntactic constructions, or speech act types is tedious and inaccurate, and adding additional coding categories is messy and error-prone. However, with increasingly accessible word-processing and database management programs, the tasks of entering, coding, and analyzing speech transcripts is becoming simple, pleasant, and reliable. With data entry facilities of the type provided in programs such as CALARSP and PEPPER, or with larger DBMS systems such as Datatrieve, ZOG, RS-1, GRITS, and INGRES, the task of adding coding categories becomes feasible and relatively straight-forward.

We have found that the use of computational processing opens up new vistas for child language research much the way that the advent of recording techniques opened up the field in the 1950's and 1960's. However, there is a major and interesting difference between these two revolutions. In the first revolution, each researcher could scrape together money for a few taperecorders and microphones, and, if s/he was lucky the budget might even accommodate a system for videotaping. In the case of the new computer technology, individual researchers may be able to purchase microcomputers for entry of their own data. However, they will often find that they cannot afford the major costs involved in main-frame processing designed to make analyses across multiple data sets. Thus, it appears that the researcher is in a position where s/he can only make full use of the new technology by engaging in some form of sharing of data, hardware, and software with other researchers. Unfortunately, the scarcity of funds for collaborative enterprises has slowed the application of the new technology to the data. Only by establishing a well-organized system for the support of data exchange and the maintenance of software analyses can this new prospect in child language research become a reality.

A glance at a fairly representative publication in child language -- the annual Stanford Papers and Reports on Child Language Development -- indicates the extent to which failure to use automated schemes for searching is slowing progress in the field. In just the 20 papers in the 1983 volume from Stanford there are three papers based on paper-and-pencil counts of numbers of occurrences of different word types in the corpus of data collected by Roger Brown. Moreover, there are three other papers based on analyses of corpora other than that of Brown that could have utilized automated search and analysis techniques. By not using such techniques, researchers are not only wasting their time, but also increasing the likelihood of error and decreasing the explicitness of their analytic technique.

3. The Formation of the System

In the summer of 1981 Dan Slobin, Willem Levelt, Susan Ervin-Tripp, and Brian MacWhinney began discussions regarding the possibility of creating an Archive for typed and hand-written transcripts at Nijmegen. The possibility of adding computerized transcripts to the database was also considered. In January of 1984, the MacArthur Foundation awarded a two-year grant to Carnegie-Mellon University for the establishment of the Child Language Data Exchange System with Brian MacWhinney and Catherine Snow as Principal Investigators. These funds provided for the entry of data into the system management of correspondence, and the convening of a meeting of the Advisory Board of the System.

During the period between January and July, the following already-computerized corpora were entered into the database at CMU.

1. The data on children gathered by Susan Ervin-Tripp and Wick Miller in the early 1960's.
2. Dan Slobin's study of English, Italian, and Turkish along with full semantic codings and English glosses.
3. Bill Hall's data on school-aged lower-class children.
4. Gordon Well's time-sampled data from British children.
5. Jan Vorster's glossed and coded data from Afrikaans.
6. Brian MacWhinney's tape-recorded-diary of his two sons' development.
7. Brian MacWhinney's and Elizabeth Bates' data on sentence production in 3, 6, and 10 year olds in English, Italian, Hungarian, and Serbo-Croatian.
8. Two lists based on Kucera and Francis (1982)--one in alphabetical order and one in order of frequency.

By Kurzweil optical scanning of printed and typed materials, we have also entered:

1. Roger Brown's data from Adam, Eve, and Sarah.
2. Jacqueline Sachs' detailed study of her daughter Naomi.
3. the transcripts appended to Lois Bloom's One Word at a Time, and
4. passages from Susan Isaac's the Intellectual Growth in Young Children.

For all of the published materials, permission to copy has been secured from the publishers and authors.

Because of their historical importance, we also plan to enter the hand-written transcripts of Bloom and the diary notes of Werner Leopold. Books by Miller, Gvozdev, Leopold, and others will be entered by scanning in the near future.

The first meeting of the Advisory Board for the System was held in Concord, Massachusetts between March 15 and March 18. The meeting was organized by Catherine Snow. The board members present were Elizabeth Bates, Ursula Bellugi, Lois Bloom, Melissa Bowerman, Robin Chapman, Eve Clark, Jane Edwards, Susan Ervin-Tripp, Paul Fletcher, Willem Levelt, Brian MacWhinney, Jon Miller, Ann Peters, Dan Slobin, and Catherine Snow. At this meeting, the Board sketched out the shape of this new System. These decisions delineated 1) the organization of the system, 2) the shape of the database, and 3) programming for the use of the database.

3.1 The Organization of the System

The System has six components: the Advisory Board, the centers, the database, programs, users, and contributors.

3.1.1 The Advisory Board

The Advisory Board is composed of child language researchers actively involved in the use of computers with transcript data. Its goal is to provide the Centers with technical advice and assistance in four areas. For each of the four areas, the Board constituted a working committee to form policies and decisions. The committees are as follows:

1. The **Traffic Control Committee** decides which corpora should be entered into the database and in what priority.

2. The **Transcription and Coding Committee** developes standards for codes on the various tiers (see below).

3. The **Programming Committee** decides which programs should be written or purchased and in what priority.

4. The **Project Development Committee** proposes major uses of the database and extensions of the System.

3.1.2 The Centers

Currently, the two centers of the System are located at Carnegie-Mellon University in Pittsburgh and the Max-Planck-Institut fur Psycholinguistik in Nijmegen, Holland. Further centers can be established at those sites that are able to dedicate resources to the maintenance and development of the System. The basic functions and shape of the system are duplicated at each center. The centers keep in correspondence by computer mail, phone, and regular mail for updating of files and task sharing.

3.1.3 Users

Membership in the System is open. However, members must agree to abide by the rules of the System, not to distribute copies of programs or files without permission, to abide by the wishes of data contributors, and to properly acknowledge the contributors and the system. Members are urged to support the progress of the System through contributions of data, programming expertise, or professional advice. The centers have made announcements regarding the system at Stanford and

Austin conferences and announcements have been placed into three journals. Later this year the Journal of Child Language will publish an article containing much of the material in the current newsletter. Further announcements will be made at Boston, SRCD, and several European meetings. Announcements have been mailed to the membership of ASHA, AERA, NECLA, ICLA, and the Stanford Child Language Forum.

Members can make use of the database by 1) requesting tape copies of files, 2) logging on to the CMU-PSY-A computer by telenet or telnet, 3) sending jobs for analysis to CMU via ARPANET, BITNET or UUCP, or 4) coming to one of the centers to log on locally. Only alternatives (1) and (4) are currently implemented. Implementation of alternatives (2) and (3) must await further funding. Transcript Analysis, the newsletter of the ChiLDES, will regularly brief members on the status of the database, the use of programs, and issues in analysis and coding.

3.1.4 Contributors

The Centers are responsible for soliciting contributions of data from members and others. This is done in accord with the suggestions of the traffic control committee. Contributors are asked to complete a contribution release form. This form gives permission to the Centers to circulate copies of the files. It also allows the contributor's to set specific limitations and conditions on the use of their data. The Centers attempt to secure users' compliance with these requests of the contributors.

3.1.5 The Database

The Centers will enter data sets into the System. Some data sets are already computerized, others must be typed in or entered by optical scanning. In such cases the correctness of the entry must be verified and some reformatting may be needed. Data sets will be entered into the system even if they are not in compliance with the new transcription standards. However, potential contributors will be encouraged to transcribe new data according to the standards set by the Transcription and Coding Committee of the Advisory Board.

3.1.6 Programs

The Centers are currently designing a series of programs for analysis and database management. These programs are being designed to maximize accessibility to all users. They are described in detail in section 3.3. Members are urged to contribute programs that they have developed.

3.2 The shape of the database

The System will store 1) files with programs for database maintenance and transcript analysis, 2) administrative files, 3) archived transcript materials as originally received from contributors, and 4) transcript files actively undergoing reformatting and the addition of new codes. This fourth type of file will be referred to as the "active database." Users will be encouraged to add a variety of data to the active database. Entire tiers of coding (see 3.22) can be added to files or corpora. Bibliographical information, documentation on usage, administrative information, and commentary and evaluation regarding data analysis and data adequacy will also be stored. The text files themselves will be modified in some cases by the text-editing facilities provided by EMACS. Additional text-editing can be done within the GRITS and HUM UNIX facilities. Either Datatrieve or INGRES will be used to add hierarchical coding to files.

The transcript files of the database will be divided into archived files and active files. Archived files are files as originally received from contributors. These files will never be modified in any way. The active database will be composed of files that have been modified to regularize formatting and coding. Within these files, there is a five-level hierarchy of file information. There will be several partitions of the database by subject matter. Within each partition, data will be grouped by **projects**. A project is a major body of data donated by a contributor or team. Within a project, the data for a single informant will constitute a **corpus**. Thus for Roger Brown's child language project there are three corpora: Adam, Eve, and Sarah. Within each corpus, there are a series of **files** with each file corresponding to a recording session. Finally within each file there may be any number of **episodes** as defined by the contributor.

Files will be received from and sent to members in standard ASCII form. The preferred form of transmission will be IBM-compatible 9-track 800 or 1600 bpi ANSII-labeled magnetic tapes. The default recordlength for tapes produced by the Centers is 2048 bytes and the default blocksize is also 2048 bytes, but other values can be accepted. Tapes are made on a VAX 750 or 780 running VMS. The VMS copy command is used and the tape drive is mounted as foreign.[2] When sending tapes to the System, it would be helpful to have tapes prepared in this same way. However, it is usually possible to read tapes that are made in other ways on other systems.

3.2.1 Headers

Each of the five higher levels of the database (partition, project, corpus, file, episode) can be notated by headers. Header information is moved up to the highest possible level. For example, information that applies to all files in a corpus is stored not as a header on each file but as a corpus header. Contributors and the system will be responsible for encoding header information. The types of information to be coded are given below in order of relative generality. This information early in this list will probably apply to whole databases whereas information late in the list will apply to episodes within files.

1. **Basic Understandings**. The system sets certain basic conditions on usage of any material. These include agreement to not distribute files and to acknowledge data sources.

2. **Permissions and Restrictions**. Contributors can set particular restrictions on the use of their data. To date no such restrictions have been set by any contributor.

3. **Warnings on the Label**. Contributors and the system may warn the user about particular limitations in the use of data from projects. For example, if an investigator paid no attention to correct transcription of speech errors, this may be noted as a project header.

4. **Human Subjects**. Headers will include information on whether informants gave release

---

[2] The exact commands issued to VMS are
$ allocate mta0:
$ mount/foreign mta0:
$ copy/log [childes...]*.* mta0:*.*
Note that "mta0:" is a name for the tape drive on the CMU system. On other VMS systems, there will be a different name for the tape drive.

for use of their data and whether pseudonyms have been used to preserve informant anonymity.

5. **History**. There should be detailed information on the history of the project. How was funding obtained? What were the goals of the project? How was data collected? What was the sampling procedure? How was transcription done? What was ignored in transcription? Were transcribers trained? Was reliability checked? Was coding done? What codes were used? Was the material computerized? How?

6. **Codes**. If the project-particular codes are being used, these should be described.

7. **Biography Files**. Where possible, extensive demographic, dialectological, and psychometric data should be provided for each informant. There should be information on topics such as age, sex, siblings, schooling, occupation, previous residences, language background of the family, religion, interests, friends, etc.

8. **Table of Contents**. There should be a brief index to the contents of the corpora.

9. **Situational Descriptions**. On the level of the file and episode there should be complete header information that should allow the user to reconstruct the situation as much as possible. Who is present? What is the layout of the room or other space? What is the social role of those present, i.e. who is usually the caregiver? What activity is in progress? Is the activity routinized and, if so, what is the nature of the routine? Is the routine occurring in its standard time, place, and personnel configuration? What objects are present which affect or assist the interaction? It will also be important to include relevant ethnographic information which would make the interaction interpretable to the user of the database (e.g. if the text is parent-child interaction before an observer, what is the culture's evaluation of behaviors such as silence, talking a lot, displaying formulaic skills, defending against challenges, etc.?).

3.2.2 Line Headers

Below the five levels that can be given headers are two further levels of data. Within files the transcript is coded in terms of a series of **utterances**. Within each utterance, data may be transcribed in terms of a number of **tiers**. In order to uniquely identify utterances and tiers a series of codes are placed at the beginning of each 80-column line in the following free format:

Delimiter	Field
`**`	utterance start symbol
`;U`	utterance number
`;L`	code(s) for the language(s) spoken in utterance
`;H`	code(s) for the hearer(s)
`;S`	code for the speaker
`;T`	tier code(s)
`;`	end of header information

For example, the line header

```
**;U79;C2;LENG;HGEO;SBIL;TTEX;
```

can be interpreted as:

Utterance Number: 79
Language: English
Hearer: George
Speaker: Bill
Tier: Text

The double asterisk is used only with the main text line (TEX tier). Before secondary tiers, the single asterisk is used. These signs are used to mark the beginning of major records and for fields under this major record for programs such as SALT and GRITS. Of the six line header codes, three will normally be omitted. These are the ;L, ;H, and ;T codes. In monolingual corpora, the language will be uniform over all files and need not be noted. In dyadic interactions the hearer is always the other participant and need not be coded. In many texts, the only line coded is the text line, and this taken as default. When there are multiple languages, multiple hearers, or embedded tiers, the ;L, ;H, and ;T codes must be used repeatedly. Thus, the line header

*;U135;LENG;LSPA;HGEO;HROS;SFER;TEXP;TPAR;

is interpreted as "utterance 135, message in both English and Spanish, spoken to George and Rosaldo by Ferdinando, coding of the parts of speech of the explanation line."

Note that this scheme for line identification does not require a fixed format. It would be best to place the asterisks in the first column for clarity. However, long lines may be continued without repeating the line identifier. The occurrence of an asterisk is sufficient to mark the end of a tier of coding and the occurrence of double asterisks marks the end of all tiers of coding for a single utterance. When searching, it is possible to exclude all material within closed delimiters such as the curly brackets that mark line headers or the square brackets that mark comments.

3.2.3 Coding within the Basic Text Line

A major aspect of the coding system is the extensive reliance on lines parallel to the basic text line to code additional comments and information. These additional lines are called "tiers". The basic line of text is called the "text tier". In the simplest case, this is the series of words the speaker produces with each word separated as in normal English text by spaces. If other tiers make reference to not just words but actually morphemes, the morphemes should be preceded by hyphens. Thus "dogs" can be coded on the text line as either dogs or dog-s. Both prefixes and suffixes are preceded by the hyphen. In order to properly distinguish prefixes from stems, morphemically-segmented text lines should have stems preceded by number signs (#) when there are prefixes. Thus, the morphemic segmentation for "undecipherable" would be un#decipher-able. When there are no prefixes, the number sign can be omitted. If a morpheme has been omitted from some obligatory context (Brown, 1973), the omission can be marked by a percent sign. Thus, the utterance that go here could be notated as: that go-%3S here. And the utterance he here with "is" missing could be notated as: he %is here. A few additional conventions for notating the text line are as follows:

x	a syllable of nondecipherable material
pcf	a phonological consistent form
.	the end of an unmarked (declarative) utterance
?	the end of a question
!	the end of an exclamation
-	the end of an incomplete utterance
	contraction
	intonational pause
<	break missing where required ("latching of turns")
>	normal ceding of a turn

Any comment-like material may be inserted into the text line at any point by placing the comment into square brackets. For example, coughing can be marked in the following way:

*:U92:SBIL: I need a [coughs] milkie.

After the left square bracket there may be a code for the comment type. For example, [P1.800] means an 1800 millisecond pause, i.e a pause of 1 second and 800 milliseconds. If a pause lasts into the minutes, this is recorded before a colon. Thus a pause of 2 minutes and 20 seconds would be [P2:20.] with no milliseconds recorded. Turn-internal pauses may be marked as [IP...] and pauses between turns can be marked as [EP...]. The total duration of episodes and sessions can be marked in the headers.

A major problem facing those who wish to search out all occurrences of a given morpheme is that the spellings of morphemes are often inconsistent. Inconsistencies are produced by spelling rules, phonological processes, dialectal variation and so on. Full standardization of the spellings of morphemes may not be desirable. Researchers often wish to enter pseudo-phonetic characterizations into the main text line to mark certain colloquialisms and stylistic features. For example, the SALT analysis program (Miller and Chapman, 1983) recognizes these spelling conventions for English:

ain't	atta(= that's a)	betcha	doctor(= Dr.)
gonna	gotta	hafta	hey
hi	huh	liketa	lookit
mhm	mister(= Mr.)	Mrs(= Mrs.)	Ms(= Ms.)
nope	oops	ok(= okay)	oughta
psst	sposta	trynta	uhhuh(= no)
wanna	whatcha	ya(= you)	yeah(= yes)
yep(= yes)	cuz(= because)		

Closely related to this problem is the introduction of variation by spelling rules and allomorphy. Thus the morpheme have occurs as have in have, but as hav in having and the morpheme wife appears as wive in wives. Our solution to this problem is to rely upon the computer to remind us about allomorphy, spelling rules and colloquial pronunciations. Connected to the search mechanism within the overall format of a system like GRITS, we will have a dictionary of allomorphs and orthomorphs. When a user decides to search for all occurrences of "wife", the system will send a message advising the user that there is an additional spelling that could be searched: wive. The same facility will inform the user about possible homonymies. It will advise the user of ways of dealing with homonymy. For example, in order to distinguish the verbal suffix -es from the plural suffix -es, the user can select an automatic look-up of the part-of-speech of the previous stem in the part-of-speech table. However, even this heuristic will not solve all cases. For example, the stem hoe in hoes could be either a noun or a verb. In order to fully distinguish homophones, a comment field can be used in the main text line. Thus, goes would be go-es[3S] and dogs would be dog-s[PL]

If material within the comment refers to a particular stretch of speech, the stretch can be marked by angle brackets. For example, speaker overlap can be indicated as follows:

*:U181:SMAR: But I <wanted to go> [overlap 182]
*:U182:SFRE: <No, you> [overlap 181] can't.

Or the stressing of a string of words can be marked in the following way:

*:U231:SMAR: Why <can't I have cake?> [stressed]

Possible alternative transcriptions can also be coded as comments

*:U919:STOM: I think he <did a little dance> [Alt/didn't take a chance]

3.2.4 Coding on Tiers

Because many researchers are interested in supplementing the basic code line with a variety of additional detailed codes that are more extensive than comments in the text line, it is important to provide a uniform way of entering this additional material. We propose to do this by developing a set of codes for tiers that are placed in relation to the main text line much in the way that parts of a musical score are placed in relation to the melodic line. For each tier recognized by the system, a set of codes will be developed on the basis of advice provided by researchers who specialize in the relevant subject area. In some areas, these codes have already undergone some standardization. In other areas, full standardization is a long way off and the best be can achieve is dissemination of possible codings. Where necessary, small working meetings will be held to settle on conventions.

In the actual files all qualifiers or codes will be marked with crosshatches and will be abbreviated as much as possible. Thus the code for "intentionality" will be $INTEN. Here, however we just give the unabbreviated qualifiers. Coding tiers that we currently recognize include:

1. **Time Marking**: Researchers who have equipment that permits full time-stamping from tape or videotape will add exact hours, minutes, and seconds on this line. Those who do not have time-stamping equipment may make use of stopwatch times or the feet or meters registered by counters on taperecorders. Entry of information on this tier is helpful, but not obligatory, since temporal relations can be judged by pegging markings on secondary tiers to words on the main text line.

2. **Morphemic Semantics**: Some standardization has been achieved in the working out a coding system for the grammatical/semantic features of individual morphemes in different languages. Comrie (1981) has published a proposal for standardization in this area and that proposal has been adopted by Dan Slobin in his forthcoming set of edited papers on the cross-linguistic study of language acquisition (Slobin, 1984). These codes serve to provide an accurate morpheme-by-morpheme glossing for non-English transcripts. In order to insure that the codes correctly match the morphemes of the text line, the morphemes of the text line and the codings of the morpheme semantics line must both be separated by hyphens. If more than one semantic code applies to a morpheme, the codes should be concatenated by slashes. Thus, PL/ACC/MASC/FIRST codes for the first declension masculine accusative plural morpheme -i in Latin. When there is overt morphological marking of agreement, this marking is to be coded at each site. In addition to the semantic codes, a set of morphophonological comment codes may be included on the morphosemantic line. These comment terms include: oblique, compound, contraction, metathesization, etc. For ASL morphophonological commenting one can use terms such as: invented, cooccurrent, fingerspelling, and pantomime. If comments refer to a single morpheme they are placed into square brackets following the morpheme to which they refer. For example, to note that the stem wive in the form wives is oblique we would have

;U1916;SKEI; Seven wive-s?

;U1916;SKEI;TMO; NUM COMMON[oblique]-PL

Some of the semantic codes of the Slobin system refer to whole clauses. These codes will not be used on the morphemic semantic tier. Rather, they will be entered on the clausal relations tier.

These morphemic codings go a long way toward representing many of the most widely

studied semantic intentions. However, it will often occur that a child uses a sentence to express one of these semantic intentions even when there is no morphemic form on the surface to code the intention. In such cases, it is necessary to insert dummy or empty morphemes on the text line. This amounts to claiming that some material has been ellipsed from the text. This same mechanism can be used to code for the semantics underlying ellipsis in adults.

3. **Structural Coding**: A large number of coding categories refer not to individual morphemes, but to high groupings such as phrases, clauses, and sentences. In order to properly relate these structural codes to the units to which they refer, one needs a way of encoding hierarchical relations into the database. The most economical way of notating syntactic structure simply uses parentheses with subscripts as in:

   (((MY)adj (DOG)noun)np ((HAS)verb (FLEAS)np)vp)s

   A more flexible and more readable system would make use of a hierarchical database structure such as DEC's Datatrieve to enter words on the text line into multiple fields in a relational database. This is the approach currently being taken by Gavin in his CALARSP system. Datatrieve provides full query-language facilities for data entry and output facilities for analysis of such hierarchical structures. For each clause or utterance, the user sees the full hierarchical structure displayed on the screen and can edit this structure in order to insert the required codes. With this facility, codings such as subject, object, topic, comment, coordinate clause, subordinate clause, foregrounded clause, backgrounded clause, appositive, and so on can be entered on any level defined by the basic principles of syntactic bracketing.

4. **Prosody**: No commitment to a system for coding prosody has been made. However, the scheme in Crystal (1969) might be adaptable for this purpose.

5. **Paralinguistics**: Here, again, Crystal (1969) presents a possible scheme. Major phrase boundaries will be coded on the text tier with the symbols [.] and [ ]. Hesitation phenomena will be coded on a separate tier. Stress and pitch modulations will be coded on the prosodic tier.

6. **Phonetics**: IPA notation can be used for this purpose. However, a conversion to ASCII codes must be devised. A working session to resolve some of these issues is currently being organized.

7. **Explanation of speaker's meaning**: This line is used to clarify unclear and missing referents in the child's speech. Some researchers refer to these explanations as "glosses." Ellipsed or omitted material may be explained on this line.

8. **Alternative Transcription**: If the transcriber is unsure of the correct transcription of a passage, an alternative transcription can be entered on this tier.

9. **Situational/Contextual Coding**: As noted earlier, major situational/contextual information should be noted in the headers to corpora, files, and episodes. Information that changes more rapidly should be coded on the situational/contextual tier. This information should allow the user to reconstruct the ongoing flow of activities. Note, however, that some of the most detailed situational changes may also be coded on the gestural and proxemic tiers.

10. **Speech Acts**: Coding on this line will utilize a set of qualifiers in angle brackets. Some of the most important speech act/interactional qualifiers are: question command, request,

invitation, prompt, suggestion, repetition, expansion, elaboration, break-down, rephrasing, completion, response, imitations, affirmative answer, negative answer, answer to yes-no question, answer to wh-question, request for repetition, compliance, denial, refusal, noncompliance, and choice amoung alternatives. A recent attempt to codify some of these categories is developed in Ninio and Wheeler (1984). Although it is far too early to reach full codification in this area, the development of an awareness of abbreviations and qualifiers will be useful.

11. **Gesture and Proxemics**: A variety of gestural and proxemic codes have been developed by workers in Los Angeles, Chicago, and elsewhere. Work on ASL notation is also relevant. A working meeting will be convened to support some standardization in this area. Among items to be coded are: gaze direction, reaching toward objects, manipulation of objects, examination of objects, postural changes, muscle tensings, distances between speakers, etc.

12. **Free Translation**: For transcripts in languages other than English, a free English translation can be provided as a utility to those not familiar with the language.

13. **Errors**: A separate tier is devoted to errors. Some qualifiers for that tier include: secondary stem overgeneralization, primary stem overgeneralization, consonant assimilation error, harmony error, selection error, sandhi error, tone error, segmentation error, superfluity, contradiction, redundancy, overanalysis, neologism, incorrect morpheme order, agreement error, affix semantic extension, stem semantic extension, blend, malapropism, anticipation, perseveration, exchange, omission, and stranding.

14. **Hesitations**: The coding of hesitations will include: retraced false start, word repetition, syllable repetition, drawling, consonant repetitition, filled pausing, and imcompletions. The types of fillers for filled pauses will be notated and the durations of pauses will be coded. Overlapping of speakers will be coded directly by comments on the text tier. Latching (additions which follow the rhythm of the partner's speech, allowing no pause between speakers) will be coded on the main text line with the < symbol. The same symbol should be repeated on this line.

15. **Comments**: General comments can be placed on the comment tier.

In order to map the relation of comment tiers to the text tier we use numbers in square brackets to mark the words on the text line where the comment starts. The codings themselves are placed in angle brackets. For example, if the phonetic line refers to words 2, 3, 4, 6, and 7 of the text line, the following coding is used.

;U812;SNED; He wanted to go to the store, Mom.
;U812;SNED;TTE; <wanted tu go> [2-4] <de stawr> [6-7]

All of the above standards for coding are -- at this point -- open to revision. The editor of this newsletter is now soliciting in-depth criticism of all of these proposals. Comments may be submitted as letters to the editor or as simply not-to-be-published comments.

3.3 Programs

The centers are currently involved in the acquisition and development of several types of computer programs for the System. These programs are designed to facilitate utilization and updating of the database. Currently available packaged programs include SALT (Systematic Analysis of Language Transcripts), OCP (Oxford Concordance Programs), and the HUM UNIX C Routines. The SALT package at the Center at CMU only runs on an Apple II and cannot be run on the mainframe. Source code is available for OCP and HUM so that they can be elaborated. In addition several Pascal utilities have been written at CMU for the System. The EMACS text editor provides many basic functions in an environment of user-definable functions. Particularly useful is the capacity for recursive calls to the editor which permits the user to edit a group of files interactively. For less interactive editing, the UNIX facilities are appropriate. Together, these programs allow the user to perform these analyses.

1. **Search.** The most powerful search function is that provided by EMACS. The search string is any UNIX regular search expression. Users will be able to search for specific words, codes with number signs, comments in square brackets, line identifiers with curly brackets or any combination of any of the above with any type of intervening material. SALT also provides highly general search facilities.

2. **Frequency tables.** The various packages all perform frequency counts by word or string types. The tables generated can be of these types:

    - **Complete frequency analyses of a text.** The numbers of occurrences of all words can be output in numerical or alphabetical order. Here words can be characterized as strings preceded by spaces.

    - **Frequency analyses by line header type.** Using the codes in the line headers, programs like OCP can generate complete alphabetical and/or numerical frequency analyses for any given speaker or combination of speakers, language or combination of languages, hearer or combination of hearers.

    - **Morpheme-based tables.** When morphemic structure has been coded on the main text line by use of hyphens, it will be possible to generate tables for only stems (items preceded by a space or a number sign), only prefixes (items followed by the number sign), or only suffixes (items preceded by hyphens).

    - **Summary statistics.** The frequency lists can be accompanied by summary statistics such as type-token ratios, total number of tokens, total number of types.

3. **Mean Length of Utterance.** The programs will be able to calculate the number of utterances, the number of morpheme tokens and then return the ratio of morphemes divided by utterances.

4. **Pause Number and Percentage.** If pausing has been coded in seconds, the entries can be counted and summed to yield total number and time of pausing.

5. **Concordances.** All of the packages can create KWIC (Key Word In Context) indices. These are essentially searches in which a specified amount of material before and after the matching string is included in an output file.

6. **Productivity Analysis.** From concordance-like data one can construct productivity analyses such as those proposed by Ingram and Dale.

7. **Dumps and Listings.** Users may often want to dump out the whole file in some new format. For example, they may wish to extract from a dyadic interaction only the lines spoken by the parent. Or they may want to reformat data so that the mother's speech is in a column on the right and the child's in a column on the right. The centers will attempt to provide general utilities for generating outputs of this type.

8. **Assisted Data Entry.** A LISP program written at CMU is now being used to automatize the process of binding coding to tiers and/or adding new tiers for coding. Datatrieve offers a way of linking such coding to a full database structure.

9. **Automatic Tagging.** Systems such as that described by Marshall (1983) and Johansson (1982) for coding part of speech in major corpora such as the Brown and LOB corpora will be implemented for the ChiLDES database.

In the long run the programs for analyzing the database will be integrated into a single environment. The design of that environment is provided by Tom Peters' GRITS. The system will allow the user to move back and forth from searching, to dictionary checking, to advice-seeking from expert programs, to statistical analysis, and printing. GRITS is currently being implemented with windows and mice on the SUN workstation. In the near future it will be moved to the new IBM machine that CMU will be deploying beginning in 1986. This machine will run UNIX and provide extensive networking facilities to and from DEC equipment.

4. Limitations and Problems

In a new undertaking of this kind it is important to consider the possibility that what appears to be a positive contribution to research might lead to negative consequences for the field. It is extremely important to maintain open access to the database for all researchers. It is also important for users to contribute their data to the system. The policies of the System must be directed toward encouraging openness and sharing. Data of all types must be represented. The over-representation of children of American college professors and the under-representation of non-mainstream children in the child language database must be corrected. Although the system cannot process audiotapes and videotapes, it should encourage researchers to make available copies of these tapes when resources permit.

Some researchers worry that having a ready source of data will take away the need to collect new data. It may be true that there will be an increase in emphasis on analysis of this huge new database. However, there will always be a need for the collection of new data. What will change with the advent of the ChiLDES is that we will now know exactly what the lacunae in our database are and why we must collect particular new corpora.

References

Brown, R. 1973. *A first language: The early stages.* Cambridge, Massachusetts: Harvard University Press.

Comrie, B. 1981. *The languages of the Soviet Union.* New York: Cambridge University Press.

Crystal, D. 1969. *Prosodic systems and intonation in English.* Cambridge: Cambridge University Press.

Johansson, Stig. 1982. *Computer corpora in English language research.* Bergen: Norwegian Computing Centre for the Humanities.

Marshall, Ian. 1983. Choice of grammatical word-class without global syntactic analysis: tagging words in the LOB corpus. *Computers and the Humanities, 17,* 139-150.

Miller, J., & Chapman, R. 1983. *SALT: Systematic Analysis of Language Transcripts, User's Manual.* University of Wisconsin-Madison.

Slobin, D. I. 1984. *Cross-linguistic studies of language development.* Hillsdale, N.J.: Lawrence Erlbaum.

Transcript Analysis Volume 2, 1985

Dept. of Psychology
Carnegie-Mellon University

2. Archiving Data Sets: Issues and Problems

Catherine E. Snow

Harvard Graduate School of Education

Cambridge, MA 02138

The dream of establishing an archive of child language transcript data has a long history, and individual efforts along such lines are not uncommon. For example, Roger Brown's original Adam, Eve, and Sarah transcripts were typed onto stencils and mimeographed in multiple copies. The extra copies have been lent to and analyzed by a wide variety of researchers -- some of them attempting to disprove the conclusions drawn from those data by Brown himself! In addition, of course, to the copies lent out or given away for use by other researchers, a master copy -- never lent and in principle never marked on -- is retained in Roger Brown's files as the ultimate historical archive.

Such storing and lending of hard copies of transcripts formed an historical precedent for the establishment of a true, comprehensive, international, crosslinguistic child language data archive, but a revolution in the basic conception of such an archive was made possible by the emergence of computers as basic tools for storage, analysis, and communication. In the traditional model, everyone took his copy of the transcript home, developed his/her own coding scheme, applied it (usually by making pencil markings directly on the transcript), wrote a paper about the results and, if very polite, sent a copy to Roger. The original database remained untouched. The nature of each individual's coding scheme and the relationship among any set of different coding schemes could never be fully plumbed.

Computer storage and exchange can transcend these limitations, and at the same time change the basic conception of an archive. Rather than primarily serving as an historical record or as a means of short-circuiting the painful, time-consuming process of transcribing for some researchers, a computer archive can become a constantly growing dataset, enriched by every user, since anyone who 'borrows' from the system undertakes at the same time to contribute to the system. This vision of archive reinterpreted as accretional exchange system inspired the planning of the ChiLDES (Child Language Data Exchange System; MacWhinney & Snow, 1984, in press; MacWhinney, 1984), which may provide a model for future systems of this type.

2.1 Why archive?

The reasons for developing a computerized exchange system for language data are immediately obvious to anyone who has produced or analyzed transcripts:

- preparation of a transcript for analysis is extremely time-consuming, and the more carefully it is done the more time-consuming it is.
- much information is latently present in any transcript so prepared.
- only a small proportion of that information is usually of interest to the original transcript-producer.
- computerization allows for the automatization of many otherwise time-consuming and error-analysis procedures.
- no single researcher has access to the range of subjects, populations, situations, or languages that would ideally be used as a basis for describing some phenomenon or testing some hypothesis. The ideally designed computer exchange system benefits both the borrower, by giving access to transcripts otherwise unobtainable, and the lender, by providing analyses, insights, and perspectives on the data beyond what a single person could achieve. However, the interests of the borrower and the lender are not always ideally served by precisely the same design aspects of the system -- and there arises the first of many conflicts that must be resolved if such a system is to be used most efficiently and exhaustively by the greatest number of people.

2.2 Conflicts

2.2.1 Entry versus analysis, storer versus user

We all have our favorite ways to format transcripts. My own preference is to put adult speech on the left and child speech on the right -- a clear reflex of my interest in the speech directed to children. Many researchers use columns with child speech on the left, but many others use a playscript format. Decisions about these matters first made years ago have gotten to seem inevitable -- though many reflect nothing more than the mechanics of the typewriters available in the late 60's, such a tab keys. The same operations might be highly inefficient for computer entry, and might also be extremely wasteful of storage space. The format which is best for the reader in terms of easy visual parsing is often hard on computer memory, whereas the kinds of formats computer experts propose are often impossible to see clearly. Furthermore, of course, the simplest way to design an archive (though not the only way -- it is possible to provide programs that translate from a single storage format to any of a variety of formats for output) is to impose a single format on all data stored -- despite the differences in the ways different researchers want to see the material laid out, and the danger that certain phenomena become harder to see in certain formats.

The need for uniformity within the system (and the associated presumption of uniformity for data placed in the system) creates another problem if the kinds of data to be included are as rich and varied as possible. Consider a system like the ChiLDES in which both transcript data and diary data will be included. A child utterance from a diary file looks very like a child utterance from a transcript file, and of course would be treated identically by any search or analysis procedure. They are, nonetheless, rather different entities with very different information values. The diary utterance was selected for recording by the researcher because it was deemed interesting, whereas the transcript utterance was recorded and entered simply because it occurred within a particular time period. Obviously, the ideal system will make room for both kinds of data -- but will also ensure somehow that the enormous differences between these kinds of data in terms of the conclusions to be drawn from them will be respected.

2.2.2 Uniformity versus special, local needs.

From the designer's point of view, the ideal system would use only the standard ASCII characters in standard ways to represent all information. Needless to say, though, the admirable uniformity thus achieved would have considerable costs in efficiency. Suppose Professor Hesitation's transcripts are especially richly supplied with information about pauses -- their length, whether they are filled, semifilled, unfilled, inhalatory, exhalatory, accompanied by manual, ocular, or labial gestures, etc. He'd want to use the special ASCII characters (! @ # $ % ^ & * +) to code his pause information. But the system as a whole wants to reserve these characters for more widely used categories of information (like morpheme boundaries), leaving Hesitation with awkward and inefficient multi-character ways of noting his categories.

2.2.3 Completeness of entry versus accuracy and readability.

In terms of analysis, the more information that is entered with the transcript, the simpler and faster any procedure will be. However, entry of much of the information at the transcript level is subject to errors of omission -- it is hard to train people to add in all the unusual bits when typing text -- and it decreases legibility of the transcript as well. Consider the following transcriptions:

1. Mary had been suddenly knocked unconscious by hitting her head on the beam.
2. Mary have-ed be-en sudden-ly knock-en un-conscious by hit-ting her head on the beam.
3. Mary have-ed be-en sudden-ly knock-en un#conscious by hit-ng(GER) her (POSADJ) head on the(DEF) beam.

If you're going to want a morpheme count on such an utterance, in order to do an MLU for instance, then the simplest way to get it is to enter morpheme boundaries as in (1). But suppose you also want to generate at some point a list of stems used: this will require distinguishing prefixes

from suffixes as in (2). Doing a search at some point for affixes of certain types will require inputting them in a standard form:

not hit-ting but hit-ing

not im-poss-ible but in#poss-able

And suppose that the planned morphological analysis will need to be able to identify certain morphological categories that do not have a one-to-one relationship to surface form. Then one will need a transcription such as that given in (3). This utterance as transcribed in (3) is now nicely prepared for an indepth analysis -- but it is almost impossible to read, it was hard to type, it required some training in linguistics to generate, and processing many utterances in this way might well produce an unacceptable high error rate.

2.3 Other Problems

The conflicts identified above are presented, not because they are insuperable, but simply to suggest that making a system adapted to one purpose may make it less adapted to another purpose. Maximizing utility and efficiency for lender and for borrower, for human user and for machine analyser, for the computer jock and the computer neophyte, may not be possible. One needs to have as a goal, rather, avoiding extremes of inefficiency, waste, and inaccessibility. Some inefficiency, waste, and inaccessibility will be inevitable.

A complex of problems for any data archive arises around communication with potential users. How can users best communicate with the system in ways which ensure ease of access but that do not overload the system personnel? How can access be assured to qualified users, and denied to potential misusers? Associated issues are how users will be trained to use the system, how publicity about its existence will be handled, how it will be ensured that the system acquires the data it needs, and how to encourage researchers to use it in the early stages, when use will require that they learn new skills.

Another set of problems that has received some attention in the establishment of the CHILDES is the general area of 'quality control'. The quality control issue arises at many levels. Is there some procedure for deciding whether data are good enough to be entered into the system? Is there some categorization of data already entered in terms of what it is good for -- what kinds of analyses it can support? How can one ensure that data will not be misused, for example by researchers totally unfamiliar with the social, cultural, or linguistic context of its collection? These are not new problems, but they become more serious and more acute as data are stored in ways which make them remote from their originator, and in ways which promote every entry to an equal status of 'real' and 'informative'. Most users will not have access to the original recordings, and will not

contemplate the possibility that some portion of the data may have been mistranscribed, misinterpreted, or misanalyzed. Therefore the administrators of the system must be responsible for maintaining a level of quality control that will guard against the consequences of propagating error.